© 18percentgrey | AdobeStock

The Dawning of a New Era in DNA Profiling

Probabilistic genotyping (PG) is a new computer technology for interpreting complicated DNA profiles that is taking crime laboratories by storm. Labs across the country are either in the process of purchasing and validating PG or planning to purchase and validate it in the near future.

Attorneys need to be prepared to litigate cases involving PG because it is already being used in court — even though guidelines on how it is to be used have not yet been written.[1] Prominent scientists have advised that it has not yet been foundationally validated for the DNA test results that labs will want to use it for most often.[2] Many have expressed concerns that the inner workings of the software programs are known only to the programmers that have developed them.

Even though PG is still evolving, one thing is certain: PG represents a paradigm shift in the way in which DNA profiles are interpreted and reported.

This article will explain why PG constitutes a new era in DNA profiling and how attorneys can prepare for its introduction in a trial.

DNA testing is increasingly being used to test for the trace amounts of DNA left by casual touch. Two particular problems have become intertwined: (1) trace DNA samples usually contain vanishingly small amounts of DNA and, (2) they also often contain biological material from multiple contributors. Add to this the complications inherent in a crime scene sample – exposure to the elements, unknown age of the biological material and the presence of inhibitory substances (like the oil from a gun or the dye from clothing) – and this confluence poses a perfect storm of uncertainty. As the number of possible contributors to a sample increases and the amount of DNA they are bringing to a mixture decreases, the results of DNA tests quickly become uninterpretable using what have been conventional approaches. That is where PG steps in. It is being used to interpret the complex DNA profiles that cannot be interpreted by a human analyst and would very likely be reported as inconclusive, if not for PG.

PG began as a tool to aid human experts performing statistical calculations on DNA profiles from mixed samples. Some PG approaches are simple enough that humans can confirm their findings with paper and pencil. But other PG programs use tens of thousands of lines of computer code and are far beyond even their programmers' ability to confirm by hand. Making a complicated situation even more difficult, generally speaking, the more complex the PG program's algorithms, the more closely they are kept secret by the PG software's manufacturers.

Trial attorneys do not have to know how the software works, but they do have to know what to do when they face it in court. Attorneys involved in criminal trials need a basic understanding of how PG differs from conventional interpretation of DNA test results. There are serious ramifications for those that they represent if computers are relied upon to make decisions about the probative value of critical evidence samples in their trials.

BY SIMON FORD AND DAN KRANE

How the Use of PG Is a Paradigm Shift in DNA Testing

PG represents a significant change in the way that complex DNA profiles are being interpreted. The paradigm of DNA testing since its earliest days has been to first attempt to exclude an individual as a possible contributor and then, if that effort fails, to determine what fraction of a pool of alternative suspects would similarly fail to be excluded. Exclusion could sometimes be as simple as finding that a suspect had an allele that was not found in an analysis of an evidence sample. PG approaches differ, however, because they intrinsically accept that test results may be incomplete. Instead, these approaches focus less on how rare a person's DNA profile is and more on how consistent the test results are with the prosecution's theory of a case relative to individual defense theories of the case.

PG fundamentally shifts the interpretative process from the in-house laboratory analyst, who performed the testing, to a computer program. The implications of this shift are particularly important because the PG software programs range from being opaque to outright impenetrable black boxes.

Before PG, if an attorney wanted to question a reported finding (for example, that an individual cannot be excluded from a complex mixture found on an item of evidence), that attorney could take an expert through the analyst's individual decisions. Guided by the analyst's contemporaneous notes, the attorney could ask the analyst to explain which empirical analyses justify any departures from the interpretation guidelines that the lab had established with its validation studies.

In contrast, when equivalent decisions are made by PG computer programs, analysts revert to vague answers such as "validation studies have shown that the software usually gives the correct answer even when its methods are at odds with those of our own laboratory" or even just "I trust the computer." The computer cannot be cross-examined and its programmers have proven to be quick to assert protection of intellectual property as a reason for not providing substantive answers to such questions.

PG is unlikely to have an immediate impact on simple DNA samples, like single-source bloodstains. Such a sample produces a DNA profile that human experts can readily agree is easy to interpret. Conventional statistical analyses are much easier for analysts to generate, understand, and explain. With simple samples there are really just three reporting options: an inclusion ("they match"), an exclusion ("they don't match"), and inconclusive ("can't tell"). In the world of standard DNA testing, exclusions are absolute. Laboratories only calculate a statistic when there is a match.

Mixed DNA profiles can be much more challenging to interpret. It is usually difficult to even determine something as fundamental as the most likely number of contributors to a mixture. When mixed samples also exhibit commonly encountered problems associated with small amounts of DNA and/or environmental exposure, they quickly become unsuitable for interpretation using conventional approaches.

For instance, an evidence sample that has been touched by multiple individuals can give rise to a very complicated test result because the DNA profiles from each of the contributors — which singly would be much easier to interpret — are layered, one on top of another. Real-world factors, like aging and exposure to the environment (degradation), can lead to random loss of parts of the DNA profile (*drop-out*). Further, samples with very small amounts of DNA are more prone to contamination within the testing laboratory leading to a *drop-in* of genetic information that was not in the evidence sample when it was collected at a crime scene. Even artifacts that are byproducts of the testing process (stutter and pull-up) can become difficult to distinguish from a signal derived from actual DNA in the tested sample.

All of these factors vastly increase the number of viable alternative explanations for the test result that was actually observed. With such a DNA profile, inclusion/exclusion is no longer black and white — gray areas can predominate where few, maybe no one, can be definitively included or excluded as a possible contributor. It may be that there are many, many shades of gray in that some people are more confidently included or excluded than others.

Attorneys, judges, and jurors understand gray areas. The real challenge for DNA experts arises when they have to make a statistical estimate that expresses how impressed the trier of fact should be that an individual cannot be excluded. If a subjective/flexible approach is used for matching, it is hard to know how many other people have profiles that would cause them to be similarly considered a possible contributor. In a poor quality, complex DNA profile, the failure to exclude an individual as a possible contributor may mean very little, and this must be communicated to the trier of fact. Not providing a statistic is not an option in the realm of DNA profiling — abundant case law and scientific sources clearly establish that inclusions must be accompanied by a statistical estimate so that a judge or jury can properly weigh the evidence.

It is easy to say that some DNA test results are not hard to interpret while others, like mixtures where drop-out may have occurred, can quickly become uninterpretable. However, it has proven difficult to objectively identify features of test results that signal that they are outside the realm where conventional approaches give reliable statistical weights. Since they do not know where to draw the line, some laboratories are already moving in the direction of using PG approaches even for test results where human experts can agree that conventional approaches would have been reliable.

Bruce Budowle, former leader of the FBI's DNA research program, observed as long ago as 2001 that "because of the successes encountered with STR typing, it was inevitable that some individuals would endeavor to type samples containing very minute amounts of DNA."[3] Indeed, poor quality, complex DNA profiles have steadily become more common as DNA profiles are increasingly generated to assist with investigations like property crimes where relatively little biological evidence is the norm. A main appeal of using PG approaches has been the promise that it frees an analyst from spending time on the interpretation and statistical weighting of complex DNA profiles from trace DNA samples in low priority property crime cases. Laboratory managers found that analysts were spending proportionally more time on these kinds of cases than on sex and violent crimes where samples typically contain more DNA, are less likely to be complex mixtures, and are often collected close to when the crime occurred.

Ironically, this means that there will be a tendency for the most challenging DNA test results to be seen in cases typically given to the least experienced attorneys. The lower the stakes for a defendant, the more likely it is that these expert systems will go unchallenged. There is a significant risk that precedents will be established in settings where neither the prosecution nor the defense will fully explore the complicated nature of PG and the implications of expert systems drawing conclusions that previously were only made by humans.

Some very large government labs are moving quickly toward picking and validating PG software packages to interpret results and make conclusions about matches and exclusions. Providers of competing PG software packages are encouraging defense attorneys in those

jurisdictions to see if their algorithms provide more defense-friendly results from the same underlying data.

How STR DNA Profiling Works

Lab reports about STR DNA profiling generally list the samples tested and provide a chart or table showing the *DNA profile* of each sample. The *DNA profile* is a list of the *alleles* (genetic markers) found at a number of *loci* (plural for *locus*, which means a position) for which information was obtained.

STR DNA tests do not produce a readout of the genetic code. Instead they type samples by determining which alleles are present at a series of different loci (often using a commercially available test kit, Identifiler®, that examines fifteen different STR loci). The loci are chosen because they are sites where human DNA tends to be particularly variable between individuals in ways that are easy to measure.

Figure 1 shows the DNA profiles of five samples — a bloodstain from a crime scene and reference samples from four possible contributors — as represented in a typical lab report. These samples were tested with an automated instrument called a *Genetic Analyzer*. This system (as well as the Identifiler® test kit) for typing DNA was developed by a company called Applied Biosystems (ABI), a subdivision of Thermo Fisher Scientific. Other, newer test kits (such as Globalfiler® also developed by ABI and Fusion 6C developed by Promega) can be used with Genetic Analyzers to look at as many as 24 different loci while earlier test kits (such as Profiler Plus® and Cofiler®) have been routinely used by crime laboratories since the late 1990s to look at a total of just 13 loci.

Identifiler® simultaneously amplifies, identifies and labels DNA segments called *STR*s (short tandem repeats) that tend to differ in length from person to person due to variations in the number of times sets of four DNA building blocks (nucleotides) are repeated. These test kits use fluorescent dyes to label the different loci so that they can be detected by a Genetic Analyzer. The Identifiler® test kit uses four dyes, blue, green, yellow and red. Genetic Analyzers determine what alleles (types) are present by measuring the length of the labeled segments of DNA. The numbers assigned to the alleles correspond to the number of repetitions in the underlying segment.

For each locus a person has two of these segments and hence two alleles, one inherited from each parent. Sometimes only one allele is detected, which is interpreted as meaning that by chance the person inherited the same allele from each parent. (See in Figure 1, e.g., suspect one's profile at locus D21S11 and suspect two's profile at locus D7S820). However, it is common to find that most of an individual's loci have two different alleles. This makes it relatively easy to determine the minimum number of contributors to a mixed sample. The simplest explanation for a sample where a locus is found to have three or four alleles is that it is a mixture of *at least* two people because it is very unusual for a single individual to have more than two alleles at a given locus.

The Identifiler® test kit gets information from one additional locus to aid in the determination of the sex of a contributor to a sample: amelogenin. Males have X and Y versions of the alleles at the amelogenin locus; females have only the X. On the basis of the results seen for the testing of the amelogenin locus alone, two of the reference profiles shown in Figure 1 appear to be from males and two appear to be from females.

The evidence sample in Figure 1 looks very much as if it could have come from a single contributor (all loci have only one or two alleles). Direct comparisons between the evidence and reference samples shown in Figure 1 allow a determination to be made regarding which suspects could or could not have been the source of the evidence sample. Suspects one, two, and three are ruled out because they have different alleles than the evidence sample at most of the tested loci. However, suspect four has exactly the same alleles at every locus — she cannot be excluded as a possible source of the evidence sample. In a case like this, the lab report will typically say that suspects one, two, and three are excluded as possible sources of the blood, and that suspect four *matches* or is included as a possible contributor.

Behind each lab report's "Table of Alleles Detected" (Figure 1) is a set of computer-generated graphs called electropherograms that display the test results. The electropherograms shown in Figure 2 display the results for evidence and four reference samples from Figure 1 for the four loci that the Identifiler® test kit labeled with the blue dye (similar electropherograms for the loci labeled with yellow, green, and red dyes are not shown). The peaks in the electropherograms indicate the presence of human DNA. The peaks on the extreme left side of the graphs represent alleles at locus D8S1179, then moving rightward at D21S11 and D7S820, and finally CSF1PO at the extreme right. The numbers under each peak are computer-generated labels indicating which allele gave rise to each peak.

The heights of peaks in electropherograms can give a clue about the relative proportions of different contributors to mixed samples. Generally, the taller a peak, the more of its corresponding allele in the original evidence sample. For instance, if an electropherogram for an evidence sample shows that the X peak at the amelogenin locus is twice as tall as the Y peak, that is a good indication that the sample is a mixture of both male and female DNA in a roughly 2:1 ratio.

Figure 1: Table of Alleles Detected — Identifiler®

Which individual is a possible source of the crime scene sample? Only one of the four individuals has a DNA profile that matches the DNA profile observed in the evidence sample.

	Evidence	Suspect 1	Suspect 2	Suspect 3	Suspect 4
D8S1179	11, 12	15, 16	13, 15	13, 15	11, 12
D21S11	29, 29	30, 30	30, 34	28, 35	29, 29
D7S820	8, 11	11, 12	11, 11	8, 11	8, 11
CSF1PO	8, 8	8, 12	8, 13	7, 8	8, 8
D3S1358	16, 17	14, 18	16, 18	15, 16	16, 17
TH01	6, 8	6, 8	6, 9	7, 9.3	6, 8
D13S317	8, 8	12, 13	8, 12	8, 11	8, 8
D16S539	9, 14	10, 11	11, 11	9, 11	9, 14
D2S1338	16, 24	16, 23	16, 24	20, 23	16, 24
D19S433	12, 16	13, 14.2	13, 15.2	13, 13	12, 16
vWA	17, 17	18, 20	12, 15	15, 18	17, 17
TPOX	8, 8	8, 11	8, 9	8, 9	8, 8
D18S51	14, 15	13, 17	16, 17	16, 17	14, 15
Amelogenin	XX	XX	XY	XY	XX
D5S818	12, 13	12, 12	13, 13	12, 13	12, 13
FGA	20, 21	19, 21	20, 21	19, 23	20, 21

After a DNA profile has been generated for an evidence sample, the current general strategy for interpretation is: First, a determination whether the DNA profile (or some part of the DNA profile) from an evidence sample is suitable for comparison with the reference DNA profiles from possible contributors to the sample. Profiles or loci not suitable for comparison are *inconclusive* and no conclusions are drawn from them. Second, genotype information from possible contributors is considered to see if they can be excluded as contributors to the loci that are suitable for comparison.

Some labs combine these first two stages and base their decision as to whether an evidence profile, or part of an evidence profile, can be interpreted on whether or not a particular reference profile matches. This is not a good practice because it reduces the objectivity of the interpretation. It makes it possible for the genotype of the subject, not the evidence sample, to drive the process.[4]

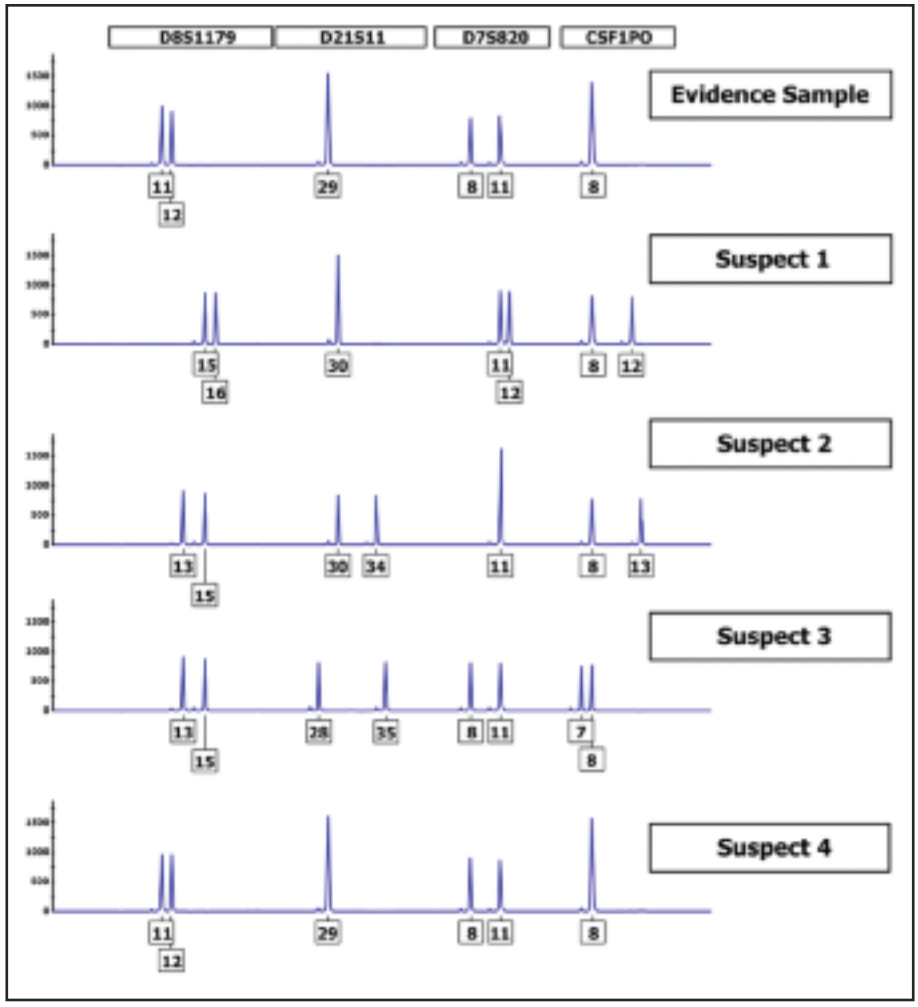
When efforts to exclude an individual (such as suspect four in Figure 1) as a possible contributor have failed, it then becomes necessary to provide the trier of fact with an estimate of the fraction of alternative suspects that would be similarly included as a possible contributor. Not all matches are equal. Complete DNA profile matches to single-source samples are exceedingly unlikely to occur by chance, but it may not be possible to exclude millions of individuals as possible contributors to the complex DNA profile from the mouthpiece of a public telephone or even from a gun that has been touched by only three or four people. The statistic should communicate to the trier of fact the robustness and rarity of the match.

Many things about the world can be measured in a way that lets people know how well a method gives results that are consistent with a ground truth, that is, information provided by direct observation, as opposed to inference. For instance, genetics theory tells us that the chance of finding a randomly chosen individual from a population that has a 15, 19 genotype at a particular DNA profiling locus is $2pq$ (two times the frequency of the first allele times the frequency of the second allele in that population). One can determine the DNA profile of 100 or 1,000 individuals from that population at that locus to see how well the method delivers relative to a ground truth — the measurable frequency of individuals who are 15, 19 at that locus. There are assumptions that underlie this genetic theory, but for a single-source DNA profile where signal is readily distinguished from noise and artifacts, it is an acceptable approximation of the ground truth and the possible points of departure are easy to understand. The approximation is also good enough that it is reasonable to extrapolate beyond what can be empirically verified at each locus by multiplying together the frequencies determined for all the loci that were tested.

It is not necessary to have formal training in the interpretation of DNA test results to recognize that some DNA profiles are easier to interpret than others. The evidence sample in Figure 2 is clearly an easy DNA profile to interpret. It appears to be a single-source sample. The heights of the peaks relative to the baseline make one confident that drop-out and drop-in are unlikely. Analysts could easily agree about what alleles were present (and absent) in the sample that was tested as well as what individuals could be excluded as a possible contributor. If a suspect was a 15, 16 at the D8S1179 locus, it would be hard to argue that the suspect's alleles were detected. In contrast, if a suspect was an 11, 12 at that locus, a statistic would be calculated to estimate what fraction of an alternative pool of suspects would also have those alleles.

The electropherogram shown in Figure 3 is more challenging to interpret than the evidence sample in Figure 2. The observation of four peaks at the D8S1179 locus and three at the D7S820 locus are good indications that the sample that was tested contains DNA from at least two contributors. As with the single-source sample in Figure 2, it is easy to be confident that drop-out and drop-in have not occurred and, therefore, what individuals can be excluded as possible contributors. But, the heights of the peaks in Figure 3 make it difficult to confidently determine what combination of alleles each of the contributors had within and between loci. A suspect with the 11 and 12 alleles at the D8S1179 locus could not be excluded as a possible contributor — but neither could someone with any of the following genotypes: 11, 11; 12, 12; 15, 15; 16, 16; 11, 15; 11, 16; 12, 15; 12, 16. Still, a statistic could easily be generated that estimated what

Figure 2: Electropherograms showing the results of Identifiler® analysis of one evidence and four reference samples at four blue loci (D8S1179, D21S11, D7S820, and CSF1PO). Which individual is a possible source of the DNA in the crime scene bloodstain? Boxes below the peaks give the name of the alleles (in number of STR repeats).



fraction of a pool of alternative suspects had one of those nine possible genotypes for the D8S1179 locus.

Some test results, like those shown in Figure 4, are simply not suitable for conventional analysis. The low peak heights suggest that suboptimal amounts of DNA were used at the start of the test. That, in turn, raises many concerns. Drop-out and/or drop-in may have occurred. The heights of low peaks are extremely variable which undermines attempts to pair alleles based on the relative heights of the peaks (unlike for the electropherogram in Figure 3). It diminishes confidence in the ability to distinguish between signal and noise and to recognize technical artifacts. Difficulty determining which alleles are present/absent translates very directly into difficulty in determining who might be excluded as a possible contributor.

Where Angels Fear to Tread

PG programs use a statistical approach known as a likelihood ratio, which is a different beast. PG approaches are not used to make predictions about the fraction of alternative suspects who have a particular genotype. Rather, they tell us how much better a set of test results supports one theory of a case relative to an alternative one. PG makes very complicated assessments as to how likely we would see a particular set of results if the prosecution's theory of a case was actually correct. It also makes a similar assessment in regard to a defense-friendly theory of the case.

That likelihood estimate is often predicated on a very large number of things such as the rate of drop-out, the ability to distinguish between signal and noise, and the extent to which a sample has been compromised by exposure to the environment. It is usually enormously impractical to determine a ground truth (e.g., create 1,000 samples for every reasonable permutation of the variables and ask what fraction of each are equally consistent with the prosecution and defense theories of the case).

Figure 4 is an example of an electropherogram that most DNA experts would agree is not suitable for interpretation or conventional approaches for statistical weighting. A likelihood ratio does not tell you one theory is right and an alternative is wrong. Instead, it provides an estimate of how confident one should be in one theory relative to another. Neither the numerator (the prosecution explanation of the DNA result) nor the denominator (generally the defense explanation of the DNA result), let alone the ratio, can be empirically compared to a knowable right answer.

All current PG approaches require very explicit statements of the prosecution and the defense hypotheses being evaluated. Even the simplest PG software currently available can only evaluate the consistency of a test result with hypotheses that explicitly assume a specific number of contributors and a specifically articulated chance that drop-out and drop-in has occurred.

As described above, determining a *minimum* number of contributors to a mixed sample is relatively easy, but there is no generally accepted means of determining the *most likely* number of contributors, let alone the *actual* number of contributors. In fact, minimum estimates have been repeatedly shown to underestimate the actual number of contributors to samples.[5]

Similarly, since the advent of STR testing in the mid 1990s, the determination of drop-out and drop-in rates has been hotly disputed by DNA profiling experts. When the probability of drop-out or drop-in is considered to be either very high (close to 1) or very low (close to 0), a PG algorithm cannot deliver helpful results, meaning that both the prosecution and defense hypotheses become very inconsistent with the test results. Those challenging DNA test results have a tendency to gravitate to extreme probabilities of drop-out or drop-in.

More complicated (and ambitious) PG software attempts to use many more parameters. These parameters distinguish between signal and noise and between peaks from real alleles and artifacts, determining the relative contributions of each contributor to a mixed sample, and the extent to which one or more contributors' DNA profile has been subject to degradation. There is a great diversity of opinion among DNA experts regarding each of these parameters and how they might or might not influence each other.

Before the advent of PG approaches, the interpretation of STR DNA test results was largely binary: peaks were either there or not there; individuals were either excluded or included as possible contributors to a sample and test results were either interpretable or inconclusive. Marginal

Figure 3: Electropherograms showing a simple mixture. Neither suspect 1 nor suspect 4 could be excluded as a possible contributor to this sample. The relative heights of the peaks are consistent with an equal mixture of DNA from those two individuals.

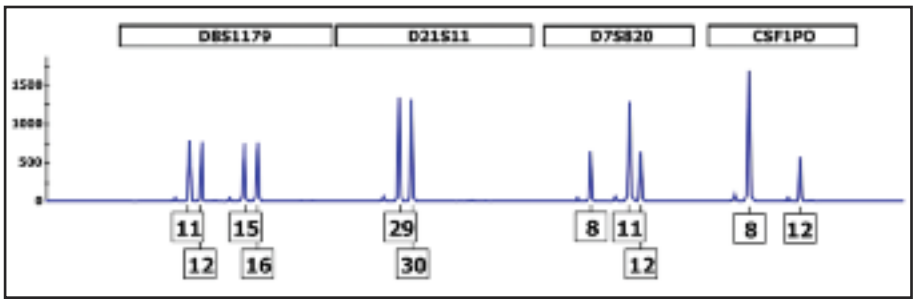
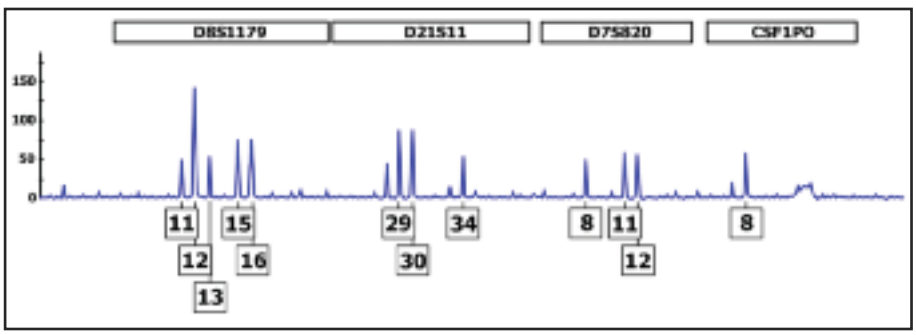


Figure 4: Low-level stochastic mixture.
All of the peaks in this electopherogram are below the stochastic threshold, indicating that drop-out is possible at any of the loci shown. Of the four suspects from Figure 2, only Suspect 4 could be considered a possible contributor without evoking drop-out. For Suspects 1 and 2 to be considered a possible contributor, there would have to be drop-out at the CSF1PO locus (for Suspect 1 a 12 allele would have to have dropped out and for Suspect 2 a 13 allele would have to have dropped out). For Suspect 3 to be considered a possible contributor, there would have to be drop-out at the D21S11 and CSF1PO loci (both a 28 and 35 allele at D21S11 would have to have dropped out as well as a 7 allele at CSF1PO).



samples ripe with stochastic effects like drop-in/drop-out and exaggerated stutter are not well-suited to black and white thinking, and those who want to interpret results obtained from suboptimal amounts of template DNA have to be willing to push back boundaries and embrace countless shades of gray.

Earlier approaches focused only on the rarity of a DNA profile in a population of alternate suspects. Most PG approaches use Bayes' Theorem to use bits of information associated with DNA test results which analysts have noticed for decades but have not incorporated into statistical weights.

Consider the possible implications of stutter — a very commonly observed technical artifact of STR DNA testing. During their validation studies, testing laboratories establish sets of rules intended to help analysts recognize and discount stutter artifacts. Both the position (one repeat unit shorter) and height (generally 15 percent or less) of a peak in an electropherogram relative to another peak at a locus are generally considered to be good indicators that a small peak could be nothing more than a stutter artifact. Until PG's emergence, little thought was given to the fact that the observation of a stutter artifact could also increase confidence that the larger peak that followed it was not noise but due to the presence of an allele in the sample being tested. With low-level samples there are often questions as to whether or not a peak is signal or noise. Considered by itself, one might feel that a particular peak was just as likely to be the result of it being signal or it being noise — in Bayesian terms, "the prior odds that it is signal are 0.5." The observation of what might be a stutter peak immediately preceding the peak in question could be used to modify those odds — perhaps to "the odds that it is signal given that it is preceded by a stutter peak are 0.75." By the same token, the failure to observe a preceding stutter peak might lessen confidence that a peak is actually due to the presence of an allele in an evidence sample. Other features (the height to width ratio of the peak, its symmetry, and the amount of background noise in areas where there should be no peaks on electropherograms) can be similarly evaluated and factored into one's confidence in each and every peak observed in an electropherogram. There is significant and ongoing debate as to how much weight should be given to the observation that a peak is preceded by a possible stutter artifact. But everyone can agree that there is a large amount of information that the traditional random match probability statistic does not capture.

Determining where to start with the prior odds for any given peak, let alone for a possible genotype for a contributor to a complex mixture where drop-out may have occurred, is not a trivial undertaking. The most sophisticated PG software considers so many features in so many different combinations and in the context of so many different prosecution and defense hypotheses that it is simply not computationally possible to evaluate even a very small fraction of them individually. In such circumstances, computer scientists sometimes use computationally intensive simulations of the data being evaluated in a method known as Markov chain Monte Carlo (MCMC). Each round of the evaluation of a sample (and there are often hundreds of thousands of rounds of evaluation that together can take hours or days to perform) begins with a different set of random number seeds that get the process of establishing prior odds started. As a result, it would be unreasonable to expect any two MCMC-driven PG analyses to give the same results for any given evidence sample (though the extent to which they are similar can be used as a clue as to how well the software is doing at finding something that approximates the best solution possible).

Courts would do well to remember the lesson of the lie detector: just because the component parts of a test are generally accepted for their individual purposes does not mean that, when they are put together for an entirely different purpose, they automatically retain that general acceptance for that different purpose. Measuring blood pressure, pulse, respiration, and skin conductivity can all be done very scientifically, but when put together do they really provide an accurate measure of assessing truthfulness?

There may be a parallel in PG. Some of the theories underlying PG are clearly well established. Bayes' Theorem goes back to 1763 and Markov published his chain theory in 1906. But is this enough to carry PG? There are many more component parts to PG and developers have not always been forthcoming in describing them. Learning from the history of the lie detector, courts should look beyond the argument that the whole is automatically scientifically valid because the components parts have been used before.

Perhaps courts should look to the broader purpose of the likelihood ratios produced by PG — to provide the lay trier of fact with accurate guidance regarding the significance that they should give to DNA evidence in the case that they are deciding. Preliminary uses of PG indicate that when different software programs are given the same data they can produce significantly different findings (as was the case in *New York v. Hillary*).[6] This is unsettling — clearly they cannot all be right. This is going to be a difficult problem for the courts to decide.

What an Attorney Needs to Do to Be Prepared

This field is evolving rapidly. In September 2016, the President's Council of Advisors on Science and Technology (PCAST) issued a report entitled "Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods" that addressed the validity of DNA testing as applied to complex mixtures (see endnote 2). PCAST made it clear that no PG approach had been foundationally validated for use on anything but two-

Bayes' Theorem takes its name from Thomas Bayes, an 18th century English philosopher, statistician, and cleric. His contribution to statistical science was to provide a mathematical framework whereby the strength of a particular explanation or conclusion can be updated to take into account additional information. In the context of a criminal trial, it can allow the forensic scientist to present the DNA evidence in such a way that it can assist the trier of fact in assessing the guilt or innocence of the defendant in the context of all the other non-DNA evidence. As applied to DNA evidence, Bayes' Theorem is usually presented as the likelihood ratio of two alternative explanations, with the prosecution hypothesis (Hp) as numerator and the defense hypothesis (Hd) as denominator. For example, in a rape case the prosecution hypothesis might be that the DNA evidence is explained by it being a mixture of biological material from the complaining witness and the defendant, whereas the defense hypothesis might be that it is explained by being biological material from the complaining witness and an unknown person. A likelihood ratio of a million would suggest that the prosecution hypothesis is the better explanation (the defendant is "included"), whereas a likelihood ratio of 1 millionth would suggest that the defense hypothesis is the better explanation (the defendant is "excluded"). A likelihood ratio of one would indicate that the DNA evidence adds nothing to the decision ("inconclusive" with regard to the defendant).

NACDL® Staff Directory

Membership Hotline 202-872-4001

Senior Resource Counsel	Vanessa Antoun	202-465-7663	vantoun@nacdl.org
Education Manager	Akvile Athanason	202-465-7630	aathanason@nacdl.org
Assistant to the Executive Director	Tatum A. Brooks	202-465-7657	tbrooks@nacdl.org
National Affairs Assistant	Shuli Carroll	202-465-7638	scarroll@nacdl.org
Deputy Executive Director	Tom Chambers	202-465-7625	tchambers@nacdl.org
Senior Editor, The Champion®	Quintin M. Chatman	202-465-7633	qchatman@nacdl.org
Membership Director	Michael Connor	202-465-7654	mconnor@nacdl.org
Resource Counsel	Jessica DaSilva	202-465-7646	jdasilva@nacdl.org
Senior Director of Public Affairs and Communications	Ivan Dominguez	202-465-7662	idominguez@nacdl.org
Junior Graphic Designer	Julian Giles	202-465-7655	jgiles@nacdl.org
Director of Public Defense Reform and Training	Bonnie Hoffman	202-465-7649	bhoffman@nacdl.org
Director of Events	Tamara Kalacevic	202-465-7641	tkalacevic@nacdl.org
Associate Executive Director for Programs, Business Services, and Technology	Gerald Lippert	202-465-7636	glippert@nacdl.org
Senior Privacy and National Security Counsel	Jumana Musa	202-465-7658	jmusa@nacdl.org
Public Affairs & Communications Assistant	Ian Nawalinski	202-465-7624	inawalinski@nacdl.org
Associate Executive Director for Policy	Kyle O'Dowd	202-465-7626	kodowd@nacdl.org
Sales and Marketing Manager	Jason Hawthorne Petty	202-465-7637	jpetty@nacdl.org
Senior Litigation Counsel	Michael Price	202-465-7615	mprice@nacdl.org
Director of Advocacy	Monica L. Reid	202-465-7660	mreid@nacdl.org
Executive Director	Norman L. Reimer	202-465-7623	nreimer@nacdl.org
Graphics Assistant	Saira Rivera	202-465-7635	srivera@nacdl.org
Member Services Assistant	Nelle Sandridge	202-465-7639	nsandridge@nacdl.org
Senior Membership and Operations Associate	Viviana Sejas	202-465-7632	vsejas@nacdl.org
Information Services Manager	Doug Shaner	202-465-7648	dshaner@nacdl.org
Public Defense Reform and Training Counsel	Renee Spence	202-465-7651	rspence@nacdl.org
Associate Executive Director for Strategic Marketing	Jessica Stepan	202-465-7629	jstepan@nacdl.org
Manager — Multimedia Production & Sales	Koichi Take	202-465-7661	ktake@nacdl.org
Counsel for Special Projects and Foundation Manager	Daniel Weir	202-465-7640	dweir@nacdl.org
Art Director	Catherine Zlomek	202-465-7634	czlomek@nacdl.org

07022018

and three-person mixtures where the lowest minor contributor gave at least 20 percent of the DNA to the total amount of DNA in the tested sample. That means that PG is not ready for prime time on the samples where labs were most inclined to use it.

In January 2017, SWGDAM (Scientific Working Group on DNA Analysis and Methods), the body responsible for generating Guidelines for the FBI that forensic DNA testing laboratories must follow to be eligible for federal funding, issued an 80-page set of new Guidelines for interpreting STR data, replacing the previous set of guidelines from 2010.[7] The 2017 guidelines have a caveat: "[t]hese guidelines may be applicable to probabilistic genotyping, next generation sequencing, and/or rapid DNA technology in a limited capacity, but are not intended for those technologies. It is anticipated that future documents will address these new technologies and methodologies."

The basic principles of PG described in this article should continue to apply for the foreseeable future. But, specific recommendations and thresholds established in authoritative documents such as the PCAST report and SWGDAM guidelines are likely to change rapidly as validation studies and refinements to PG software occur.

There is no requirement that a lab state in its reports that it used PG. It is not going to be used in all cases, but if a report includes three- and four-person mixtures, then there is a good chance that PG was used in some capacity — particularly if the results are reported in a likelihood ratio format. If the report contains results with exclusions with weights, then it is a safe bet that the laboratory used PG.

Defense attorneys should always request all underlying scientific records in any DNA case. The records of PG will usually consist of print documents and electronic files. The documentation will contain the output from the PG software, including lengthy listings of inferred genotypes with their weights. The electronic data may include files that can only be opened by proprietary software as well as files in more common formats such as PDFs. In requesting discovery of PG in a particular case, it is important to request that records for all PG runs be provided. The choice of prosecution and defense hypotheses is critical with regard to PG, as are the input parameters for the analysis (for example, the assumption of the number of contributors or the minimum peak height threshold used). The lab analyst may have taken several shots at running

the PG software, using different hypotheses and assumptions, but may have reported only the likelihood ratios that fit in with his or her expectation of the case. Other runs may include alternative hypotheses that are more favorable to a particular party's position. In some cases, a human analyst may have had to override some of the data, such as excluding a problematic locus, to obtain the reported findings. Careful review of the complete paper and electronic output from the PG analysis can uncover this information.

The paper and electronic output from PG is extremely technical and difficult to read. The output will contain the parameters used and information about the version of the PG software used in the specific case. This is where an expert is essential. PG spans several disciplines, including molecular biology, biostatistics, and computer science. It is important that any expert has familiarity with PG methods. In some cases it may be necessary to rely on an expert who has access to and working familiarity with the PG software in order to run alternative hypotheses or assumptions.

In addition to the case-specific records, it is important to obtain discovery of foundational documentation such as the lab's standard procedures, validation research, training materials, competency studies, and proficiency tests. Recent cases have focused on disclosure of the underlying source code for PG software programs, many of which are black boxes. When the source code of a black box system is disclosed, the box would be open to independent scrutiny. The main justification for maintaining black box software is the protection of intellectual property. Courts will need to decide whether the commercial interests outweigh the right of defendants to fully examine the evidence against them.

It is important to be attentive to the version and features of the PG software used in a particular case and to ensure that they are the same that was used in the testing laboratory's validation of the PG software. At present, new versions of the most sophisticated PG software seem to be released at a very much faster pace than those for more established programs and operating systems that reside on an office computer. The danger here is that a software manufacturer may make a significant change to a program that is not easy to know about. Even subtle changes can have dramatic effects on the likelihood ratios that are generated. SWGDAM's 2015 Guidelines for the Validation of Probabilistic Genotyping Systems (see endnote 1) state that "significant change(s) to the software, defined as that which may impact interpretation or the analytical process, shall require validation prior to implementation." This means that even incremental versions of PG software should require a lab to re-do its validation studies.

Conclusion

The introduction of PG is making life much more technical and complex for judges and trial attorneys. At the same time that they are struggling to understand what PG is and how it is being used, PG technology itself is a moving target that continues to evolve.

In a highly competitive business, PG developers are protecting their PG software with assertions of intellectual property rights and aggressively vying for market share.

Attorneys who have specialized in DNA cases by grappling with both molecular biology and statistics will now have to learn to talk about computer science, or at least ask questions about it. They must ask how PG was used in their particular case. In addition, they must assure themselves that the hypotheses and assumptions run through the PG program were appropriate to their case. They must ask for every PG analysis that was run so they can be sure that nothing was missed or misrepresented. Attorneys will have to be vigilant about PG software because new versions could change the playing field substantially but be admitted as evidence with no notice. They will need to know where to find consulting experts sufficiently versed in PG to help.

They should never just trust the computer.

Notes

1. The Scientific Working Group on DNA Analysis and Methods (SWGDAM) has generated *Guidelines for the Validation of Probabilistic Genotyping Systems* (https://media.wix.com/ugd/4344b0_22776006b67c4a32a5ffc04fe3b56515.pdf), but a generally accepted means of describing results has not yet emerged.

2. Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature -Comparison Methods (2016), Executive Office of the President, President's Council of Advisors on Science and Technology (https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf).

3. B. Budowle et al., *Low Copy Number — Consideration and Caution*, Proc. 12th International Symposium on Human Identification (2001).

4. D.E. Krane et al., *Sequential Unmasking: A Means of Minimizing Observer Effects in Forensic DNA Interpretation*, 53 J. Forensic Sci. 1006 (2008).

5. D. Paoletti, T. Doom, C. Krane, M. Raymer & D.E. Krane, *Empirical Analysis of the STR Profiles Resulting from Conceptual Mixtures*, 50 J. Forensic Sci. 1361 (2005).

6. See the discussion of *New York v. Hillary* in Jessica Goldthwaite et al., *Mixing It Up: Legal Challenges to Probabilistic Genotyping Programs for DNA Mixture Analysis*, The Champion, May 2018 at 12.

7. SWGDAM Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories (approved Jan. 12, 2017), (https://media.wix.com/ugd/4344b0_2a08f65be531488caa8037ed55baf23d.pdf). ■

About the Authors

Simon Ford, Ph.D., was trained in molecular biology and biochemistry. He is President of Lexigen Science and Law Consultants, a firm specializing in providing advice to lawyers about genetic evidence. He consults on DNA cases and has conducted workshops for a number of agencies on computer analysis of STR test results.

Simon Ford
Lexigen
San Francisco, California
415-865-0600
Email sford@lexigen.com

Dan Krane, Ph.D., is a Professor in the Department of Biological Sciences at Wright State University. He helped develop Genophiler™ and is founder and president of Forensic Bioinformatic Services, Inc. A leading authority on forensic DNA evidence, he has testified as an expert witness in over 50 cases.

Dan Krane
Wright State University
Dayton, Ohio
937-426-9270
Email dan.krane@wright.edu